Monte Carlo Estimates of the Validity of Four Relevant Question Polygraph Examinations

David C. Raskin[1], Charles Honts[2], Raymond Nelson[3] and Mark Handler[4]

Abstract

Monte Carlo methods were used to calculate the distributions of grand total numerical scores of event-specific, single-issue polygraph examinations with four relevant question (RQ) test formats, such as the Air Force Modified General Question Technique (AFMGQT) and a similar format developed by researchers at the University of Utah (Utah CQT). Mean and variance seeds for total scores were calculated using the subtotal mean and variance estimates from a laboratory sample of 100 event-specific exams and from 100 confirmed field exams conducted using a three question event-specific format that included both primary and secondary RQs. Parameters included correct, incorrect, and inconclusive results of guilty and innocent cases, and the unweighted decision accuracy. Positive predictive value and negative predictive value were calculated using a base rate of .5. Detection efficiency coefficients were calculated for a single measure of effect that encompasses correct, incorrect, and inconclusive results with both guilty and innocent cases. Results indicated that the accuracy of four RQ event specific examinations evaluated with grand total scores and two-stage decision rules equaled or exceeded accuracy of other validated comparison question techniques. Unweighted accuracy was .92 to .94 for three test charts, and converged towards the .95 confidence level predicted by the probability cutscores (.05 / .05) when five test charts were used. Total scores produced equally low error rates on guilty and innocent individuals, whereas decisions based on question subtotals (spot scoring) produced high rates of false positive errors.

Keywords: *polygraph, CQT, MGQT, multi-facet exams, grand total, test data analysis.*

Monte Carlo Estimates of the Validity of Four Relevant Question Polygraph Examinations

Polygraph examiners sometimes conduct event-specific or single-issue examinations using the four-question Air Force Modified General Question Technique (AFMGQT; Department of Defense, 2006a) or a similar four-question technique developed by researchers at the University of Utah (Raskin & Honts, 2002; Raskin & Kircher, 2014). Some field examiners prefer these formats because they permit the use of questions that address different facets of a single known (or alleged) event or context. Both formats may include primary relevant questions (RQs) that address direct involvement and secondary RQs that address indirect involvement. Secondary RQs may cover participation, evidence, knowledge, or factual details regarding the incident.

[1]University of Utah
[2]Boise University and Honts, Handler and Hartwig, LLC
[3]Lafayette Instrument Company
[4]Honts, Handler and Hartwig, LLC

Acknowledgements

The views expressed in this article are solely those of the authors, and do not necessarily represent those of the University of Utah, Boise University, Honts, Handler and Hartwig, LLC or the Lafayette Instrument Company

The question sequence for one version of the AFMGQT is: Neutral, Sacrifice Relevant, Comparison-1, Relevant-1, Relevant-2, Comparison-2, Relevant-3, Relevant-4, Comparison-3 (Department of Defense 2006b). The Utah four-question CQT uses a similar question sequence: Introductory, Sacrifice Relevant, Neutral-1, Comparison-1, Relevant-1, Relevant-2, Comparison-2, Relevant-3, Relevant-4, Comparison-3, Neutral-2 (Raskin & Honts, 2002). The RQs, comparison questions (CQs), and neutral questions (N) are rotated independently for each presentation of the question sequence; only comparison and relevant stimuli are used for quantitative analysis. Additional Ns may be inserted as needed (Department of Defense, 2006a) and are inconsequential when scoring either format. The question sequences of the AFMGQT and Utah four-question CQT formats are so similar that differences between them have been described as potentially meaningless when similarly employed and evaluated (American Polygraph Association (APA), 2011) This project explored the use of test formats with four RQs in event-specific diagnostic testing contexts, such as those used during criminal investigations and evidentiary proceedings[1].

APA (2011) reported the results of AFMGQT studies containing two to four RQs using the Empirical Scoring System (ESS; Nelson *et al.,* 2011) with decisions based on subtotals for each RQ. The mean unweighted accuracy of decisions was .88, with a mean unweighted inconclusive rate of .17. The mean false positive (FP) rate was .11 and the mean false negative (FN) rate was .09. APA also reported the results of AFMGQT examinations scored with the Federal 7-position scoring method. The mean accuracy rate of decisions was .82, with a mean inconclusive rate of .20. See Appendix A for additional information. However, the APA report included no study of grand total and two-stage rules with the AFMGQT or the Utah CQT.

Decisions using the grand total involve the aggregation of all scores into a single numerical total that is compared to a cutscore or reference distribution to reach a categorical and/ or probabilistic conclusion. Use of subtotal scores involves the aggregation and comparison of test scores for individual questions with a cutscore or reference distribution. Despite their similarities in structure, only the Utah CQT, makes use of the grand total score. The grand total approach assumes that the response variance for questions regarding a single known or alleged incident is non-independent. It posits that various facets of the incident and multiple physiological measures obtained from the same examinee are correlated and share sources of variance. In contrast, historic decision policies for the AFMGQT use question subtotals under the assumption that questions that address multiple facets of an incident produce independent response variance. Scientific studies and have shown that attempts to interpret independent response variance at the level of the individual questions results in substantial or dramatic increases in false-positive errors for individual RQs (Podlesny & Truslow, 1993; Raskin & Kircher, 2014; Senter, 2003; Senter & Dollins, 2003).

The decision rules based on subtotals have taken the form of "any or all," meaning that a sufficiently negative subtotal score for *any* RQ is a positive test result (DI), whereas positive subtotals are required for every RQ for a negative test result (NDI). These decision rules introduce the *multiplicity* or *multiple testing problem*, for which statistical correction procedures have been developed to manage the potential compounding of errors. Previous research has shown that using question subtotals as the basis for test outcomes may inflate errors with criterion innocent subjects and decrease test specificity.

[1]Both formats may also be used in multi-issue contexts, though that usage is not the focus of this project.

Traditional AFMGQT cutting scores do not address these inflated false-positive errors and loss of specificity. Studies on the AFMGQT using the ESS (Nelson *et al.,* 2011) that used common statistical corrections (Abdi, 2007) to attempt to control the potential for increased false positive andclassifications have shown improved mean accuracy rates compared to the traditional cutscores.

Senter (2003) reported decision rules that use the grand total or subtotals ("spot" scoring) may provide different advantages, and their choice may depend on the testing goals. These traditional MGQT decision rules were developed on the basis of experience, as opposed to being statistically derived. Published descriptions of the reference distributions of scores from guilty and innocent persons using the four relevant-question MGQT format have not been widely available. Previous studies were conducted with traditional integer cutting scores for which the sensitivity, specificity, and probability values for grand total and sub-total cutting scores were unknown and possibly sub-optimal. Raskin and Kircher (2014) used various computer algorithms to demonstrate that total scores produce the highest accuracy of decisions, especially when combined with a 2-stage decision rule (Senter & Dollins, 2004). Their analyses showed that the use of subtotals *more than tripled* the false positive rate.

The present study investigated the range of decision accuracy and errors using grand-total and two-stage cutting scores derived from distributional parameters calculated from archival samples of event-specific diagnostic examinations scored with Utah 7-position and ESS methods. Reference distributions may allow better selection of cutting scores based on statistical estimates of expected error rates, with consideration for a determined tolerance or payoff matrix for false positive and/or false negative errors.

Methods and Results

Monte Carlo methods (Carsey & Harden, 2014; Metropolis, 1987) were used to calculate statistical reference distributions and criterion accuracy estimates for event-specific diagnostic examinations with four RQs. We first used non-parametric bootstrapping to calculate the mean and variance parameters of the distributions of guilty and innocent grand-total scores for both the Utah Numerical Scoring System (Bell *et al.*, 1999) and the ESS (Nelson, *et al.,* 2011). A parametric bootstrap[2] was then used to calculate statistical confidence intervals for several measures of accuracy of four-question event-specific examinations, including test sensitivity (true-positive rate), test specificity (true-negative rate), false-negatives, false-positives, and inconclusive results.

[2] Bootstrapping (bootstrap resampling) involves resampling data with replacement as a method of calculating statistics of interest. Bootstrapping is useful when there is no exact formula to calculate a statistic, calculations are complex, and data may not conform to required assumptions for parametric calculations. Non-parametric bootstrapping involves constructing a distribution of sampling distributions by resampling individual data points from an original data set with few assumptions regarding linearity and distributional shape. Parametric bootstrapping involves the construction of a distribution of sampling distributions by resampling data points from a statistical distribution with parameters determined by our present knowledge about the investigation context and related data. Bootstrapping and Monte Carlo methods are useful to calculate complex statistics and effects of interest but they do not satisfy the need for representative data or information to describe the real world population. In the context of polygraph testing, Monte Carlo methods can estimate the margins of uncertainty surrounding various aspects of criterion accuracy, including sensitivity, specificity, false positive, and false negative errors. Additional information about model effectiveness can be gained by comparing the computational results with data from actual experiments.

In addition, confidence intervals were calculated for the unweighted inconclusive rate, unweighted decision accuracy, accuracy excluding inconclusive results, and the detection efficiency coefficient[3] as a single measure of effect that encompasses correct, incorrect, and inconclusive results with guilty and innocent cases (Kircher, Horowitz & Raskin, 1988).

Utah 7-Point Data

Total scores for each RQ were obtained from a laboratory mock crime study of the Utah Comparison Question Test (Kircher & Raskin, 1988). Participants were 100 persons recruited from the general community though temporary help-wanted advertisements. They were paid for their participation and were offered a monetary bonus if they produced a truthful outcome in their polygraph examination. The examinations contained three RQs; two of the questions were direct accusatory questions ("Did you take that ring?" and "Did you take that ring from the desk?") and the third was an evidence-connecting question ("Do you have that ring with you now?"). Thus, the test was a multiple-facet test. The ordinal position of RQs was fixed for all repetitions of the questions. Numerical scores were assigned by the experimenter and another psychophysiologist who were both blind to the guilt status of the participants.

Initial Tests of the Utah Data. We initially determined if the direct accusatory RQs and the evidence-connecting RQ produced different average total scores. The data from the original examiner and the blind evaluator were subjected to a 2 (Guilty, Innocent) X 2 (Original Examiner, Blind Evaluator) X 3 (Questions) analysis of variance (ANOVA). Evaluators and Questions were treated as repeated-measures factors, and Guilt was a between-subjects factor. Since Mauchly's Test of Sphericity (Girden, 1992; Mauchly, 1940) was not significant, the results were reviewed without adjustment. As expected, the analysis revealed a large effect of Guilt, $F(1, 98) = 145.8$, $p < .001$, $\eta p^2 = 0.60$. The estimated marginal means for Innocent subjects was 3.3 (se = 0.33), and for Guilty subjects -2.3 (se = 0.33). There was also a significant main effect for Questions, $F(2, 196) = 4.51$, $p = .012$, $\eta p^2 = 0.04$. The means and standard errors for the three RQs were 1.1 (.32), 0.2 (.31) and 0.1 (.29), indicating that RQs in the first position produced more positive scores than RQs in the second and third positions, regardless of the subject's guilt status. Although a position effect was revealed by this analysis, it failed to indicate that the evidence-connecting RQ (R3) functioned in a manner different from the direct accusatory RQs. Pairwise comparison tests with the Šidák adjustment (Abdi, 2007; alpha = .05) indicated that R1 was different from R3 and approached significance with R2, $p = 0.055$. R2 and R3 were not significantly different. The main source of the Questions effect was the bias toward positive scores in the first relevant position. These results support the practice of rotating RQs to counterbalance the position effect across three charts.

Nonparametric Bootstrap of the Utah Data. Scores from the first three charts collected in Kircher and Raskin (1988) were used for the non-parametric bootstrap (Efron, 1981) with a smoothing procedure (Silverman & Young, 1987).

[3]Detection efficiency coefficients were described by Kircher, Horowitz, and Raskin (1988). This is the correlation coefficient calculated between the guilty status of the sample cases (coded as guilty = -1, innocent = 1) and the test result (coded as deceptive = -1, inconclusive = 0 and truthful = 1). It is useful for meta-analytic research as a single metric that is sensitive to correct, incorrect, and inconclusive results with both guilty and innocent cases. This use of the correlation assumes a linear order to the correct, inconclusive, and incorrect results by positing that inconclusive results are less optimal than correct results and less problematic than incorrect results.

The bootstrap data were used to generate reference distributions for the four-question grand total scores after rounding the four-question means and standard deviations to the nearest integer (see Appendix B). The mean of the generated reference distribution for guilty cases in the sample space was -10.7 (SD = 9.2) and mean for the innocent cases was 13.3 (SD = 10.4). The reference distributions for grand total scores were decomposed to subtotal parameter estimates by dividing both the mean and variance by the number of questions (4) and calculating the subtotal standard deviation as the square root of the subtotal variance. Resulting subtotal distributions for 7-position scores were the following: guilty subtotal mean = -2.8 (SD = 4.5), innocent subtotal mean = 3.2 (SD = 5.0). In a subsequent step, subtotal scores for the guilty and innocent cases were resampled from a standard normal distribution characterized by the Utah 7-position subtotal parameter estimates.

Empirical Scoring System

Reference distributions for ESS scores of four RQ event-specific exams were computed using the same procedures as above. A non-parametric bootstrap was used to resample the question total scores of the data from Nelson, Krapohl, and Handler (2008). These data consisted of 100 examinations conducted using the Federal Zone Comparison Technique and subsequently submitted to the United States Department of Defense confirmed-case archive. All of the examinations were field exams conducted by federal and local law enforcement agencies in the investigation of a variety of crimes, including drugs, physical assault, sex crimes, embezzlement, theft, and burglary. Examinees were adult male and female criminal suspects.

One-half the cases were confirmed as deceptive, half were confirmed as truthful; little additional information was available regarding the examinees. Confirmation was achieved via a combination of examinee confession, confession from another suspect, and extra-polygraphic evidence, though it was not known how the cases were selected for the confirmed case archive. Results from the original examiners did not always agree with the confirmation status.

The Federal Zone Comparison Technique two primary RQs (e.g., “Did you steal that Mustang?”, and “Did you steal that Mustang from the parking lot?” or “Are you the person who stole that Mustang?”) followed by a secondary RQ (e.g., “Did you help steal that Mustang?”, “Do you know how that car was disposed of?”,“Do you know who stole that Mustang?”, or “Did you plan with anyone to steal that Mustang?”). In this format, RQs typically are not rotated with each repetition of the test question sequence (Department of Defense, 2006a).

Seven student evaluators scored the 100 sample cases using ESS. The evaluators were adult males and females employed in law enforcement who were in the eighth week of a basic polygraph training program accredited by the American Polygraph Association. These evaluators had received previous instruction in the U.S. Federal 7-point and 3-point scoring methods (Department of Defense, 2006b) and had received instruction on the ESS prior to the scoring task. They were asked to refrain from formulating a conclusion regarding the deceptive or truthful status of the case while completing the numerical scoring task.

Initial Tests of the ESS Data. We initially determined if the direct accusatory RQs and the evidence-connecting RQ produced different average total scores. The data from the seven blind evaluators were subjected to a 2 (Guilty, Innocent) X 7 (Evaluators) X 3 (Questions) ANOVA. Evaluator and Questions were treated as repeated-measures factors, and Guilt was a between-subjects factor. Mauchly’s Test of Sphericity was not significant for Questions, but was significant for the effects involving Evaluators. Therefore, the degrees of freedom for Evaluators effects were adjusted by the Greenhouse-Geisser method (Abdi, 2010).

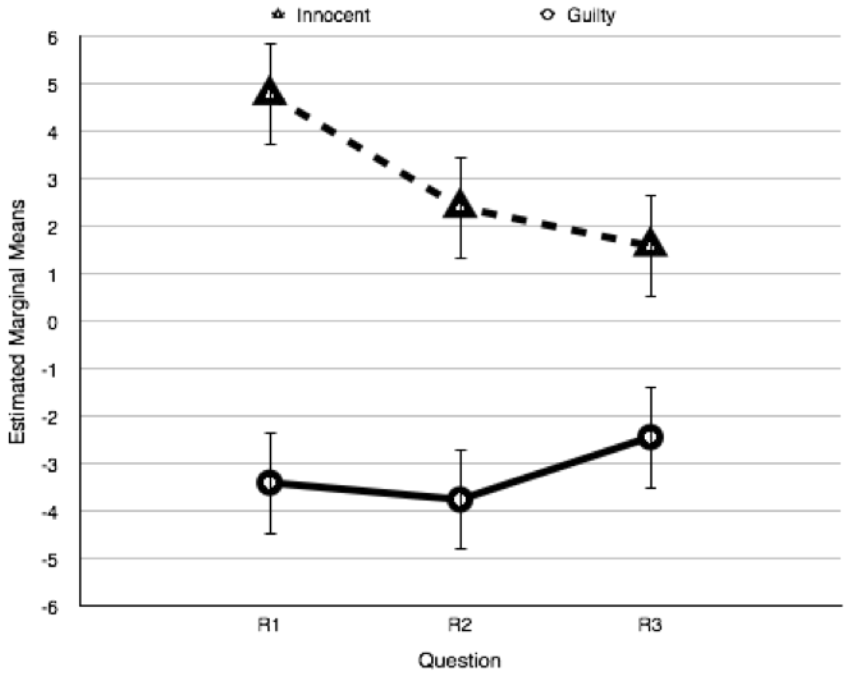
As expected, the analysis revealed a significant and large effect of Guilt, $F(1, 98) = 133.9$, $p < .001$, $\eta p2 = 0.58$. The estimated marginal means for Innocent subjects was 2.9 ($se = 0.37$) and for Guilty subjects -3.2($se = 0.37$). There was also a significant main

effect for Questions, $F(2, 196) = 5.32$, $p = .006$, $\eta p2 = 0.051$. However, that main effect had to be examined in the context of a significant Questions X Guilt interaction, $F(2, 196) = 10.84$, $p < .001$, $\eta p2 = 0.1$. Means illustrating that effect are shown in Figure 1.

Examination of Figure 1 shows that the interaction effect was due to primarily to the very strong positive score at R1 for Innocent subjects. As with the Utah scores, the ESS scores showed a position effect but failed to indicate that the evidence-connecting RQ (R3) functioned in a manner different from the direct accusatory RQs. The main source of this effect was the bias toward more positive scores in the first relevant position that occurred only with Innocent subjects. As with the Utah scores, the ESS results support the practice of rotating RQs to counterbalance the position effect across three charts. There was a number of other significant effects from the analysis of the ESS scores unrelated to the questions addressed in this paper.

One interesting difference between the Federal and Utah CQT formats is that the Federal ZCT compares the first RQ to the stronger of the preceding or subsequent CQ. In contrast, all RQs for the Utah CQT format are scored only to the preceding CQ unless that CQ is uninterpretable due to a physical movement or other artifact. The practical results of this may be a tendency for a more positive value in the numerical score of the first RQ of the Federal ZCT format than for the Utah format. This may be the source of the effect observed for R1 with innocent cases, as this procedure is most likely to affect innocent cases. Because most differences were not significant and the Monte Carlo design seeded by aggregated scores, it is unlikely that this difference had any substantial affect on the remainder of this analysis.

Figure 1. Estimated marginal means for the significant Question by Guilt interaction with ESS scores.



Nonparametric Bootstrap of the ESS Data. The mean re-sampled total scores for four RQs were -12.2 (SD = 10.2) for Guilty examinees and 11.1 (SD = 9.0) for Innocent examinees. These parameters were rounded to the nearest integer to calculate the reference table shown in Appendix C. Parameters for the reference distributions of ESS total scores for examinations with four RQs were decomposed to a reference distribution of subtotal scores for Guilty and Innocent examinees by dividing the mean and variance by the number of RQs and calculating the subtotal standard deviation from the subtotal variance. The parameters for subtotal distributions of ESS scores for Guilty examinees were mean = -3.0 (SD = 5.0) and for Innocent examinees mean = 2.8 (SD = 4.5). Subtotal scores for guilty and innocent cases were subsequently resampled from a standard normal distribution characterized by the ESS subtotal parameter estimates.

Monte Carlo Analyses

The Monte Carlo model was a parametric bootstrap of individual questions resampled from two standard normal distributions with mean and variance from the previously described nonparametric bootstrap of the guilty and innocent question scores. Use of a parametric bootstrap at this stage represents an explicit assumption that the scores are normally distributed. Two versions of the parametric bootstrap were used, one to estimate accuracy of four-RQ polygraphs with the Utah 7-point scoring method, and a second version to estimate accuracy with ESS scores.

The Monte Carlo space, analogous to a sample space, consisted of N = 100 cases for which the criterion state of each was set by comparing a random number to a base rate of .5. Because statistical theory and previous research concur with the present finding that subtotal scores from primary and secondary RQs do not vary independently, the criterion state of subtotal scores was set uniformly for each examination in the Monte Carlo space. In this way, the base rate of deception for the cases in each iteration of the Monte Carlo space was approximately .5 and converged to .5 upon numerous iterations of the Monte Carlo space. For each case in the Monte Carlo space, random numbers were standardized to the guilty or innocent question total parameter estimates according to the case status. Thus, subtotal scores within each case had a shared guilt status and shared variance from a common statistical distribution, but were generated independently. Individual question total scores were summed for a grand total score for each case in the Monte Carlo space.

In-sample accuracy and out of sample error estimation. Means and standard errors were calculated for several aspects of decision accuracy, including sensitivity, specificity, false negative errors, false positive errors, positive predictive value (PPV), negative predictive value (NPV), unweighted decision accuracy, unweighted inconclusive rate, and the detection efficiency coefficient. The out-of-sample error rate (generalization error) was estimated using 95% confidence intervals calculated as the range from the .025 to .975 quantiles of the distribution of results from 1000 iterations of the Monte Carlo model.

For Utah 7-position scores, accuracy was calculated using probability cutting scores of 0.05 and 0.05 for deceptive and truthful classifications, respectively. These probability values were selected to constrain false positive and negative error rates to a desired level and correspond to integer cutting scores of -4 and +4 for deception and truth-telling. Results were calculated for Utah 7-position scores using the grand total rule (GTR) and also using the Senter (2003) two-stage rules (TSR)(see Appendix D for procedural information regarding two-stage decision rules). To correct for the inflation of errors that results from multiplicity effects, Bonferroni correction (Abdi, 2007) was applied to the probability cutting score when using sub-total scores. The integer cutting score for sub-total scores was -10, for a probability cutting score of .0125 when using the sub-total scores.

For ESS scores, results were calculated using the TSR, including means, standard errors, and the 95% confidence interval between the .025 and .975 quantiles. Probability cutting scores were the same as

those for the Utah scores, 0.05 and 0.05 for deception and truth-telling. These probability cutting scores correspond to integer cutting scores of -4 and +5. The cutting score was -10 when using sub-total scores, for a probability of .0125 when using the Bonferroni correction with the desired probability cutting score.

To allow a more direct comparison of these data with previously reported results using ESS subtotal scores with the AFMGQT, ESS results were also calculated using the four-question format with two-stage decision rules when cutting probabilities were set at .05 and .10 for deceptive and truthful classifications. Probability cutting scores for .05 and .10 correspond to integer cutting scores of -4 and +2 for grand total scores, and a statistically corrected integer cutting score of -10 when using the sub-total scores.

Initial tests of the Monte Carlo model. To evaluate the stability of the Monte Carlo model to estimate outcomes that are similar to other methods, the detection efficiency coefficient was calculated for the raw Utah 7-position numerical scores from Kircher and Raskin (1988). The detection efficiency coefficient for the raw numerical scores from Kircher and Raskin was r = .84. This coefficient is similar to other reported detection efficiency coefficients using similar procedures, including that of Raskin and Hare (1978) r = .87 using psychopathic prison inmate subjects, and Rovner, Raskin and Kircher (1979) r = .87 using community subjects, as reported in Kircher, Horowitz, and Raskin (1988). The detection efficiency coefficient and 95% confidence interval was then calculated to compare the Model Carlo model with previously reported coefficients. The coefficient and standard error from the Monte Carlo model was r = .85 (se = .05) with a 95% confidence interval from .76 to .93 for event-specific exams with three RQs. These data suggest the Monte Carlo model was capable of providing estimates of detection efficiency similar to those from other methods of investigation.

Monte Carlo Results

Table 1 shows the means, standard errors, and 95% confidence ranges for criterion accuracy of the four-question format with Utah 7-position data and ESS data. Accuracy estimates for the four-question format were high for all scoring models, with mean unweighted accuracy over .90, except when subtotal scoring rules were used without statistical corrections to the desired probability cutscores. However, the statistical confidence ranges provide a more cautious and satisfactory estimate of the range for potential generalization error that can be expected upon the application of these testing methods with other sampling data. In particular, the lower limit of sensitivity, specificity, unweighted accuracy, PPV, and NPV may be most informative, along with the upper limit for false negative errors, false positive errors, and inconclusive results.

For unweighted accuracy, the lower limit of the 95% confidence range exceeded .86 for all models and exceeded .88 for all models when probability cutting scores were set at 0.05 and 0.05. Although no test of statistical significance was conducted, inspection of the confidence intervals indicates that a significant difference between these scoring models is unlikely. Similarity of the results obtained with the Utah 7-position and ESS scoring methods is not surprising. The ESS is based on feature development completed at the University of Utah and can be thought of as a generalization and simplification of the Utah protocol. Both methods are based on features developed at the University of Utah, and the commonality of features may be more important than the computational procedures in this context. It is of interest to note that unlike the results reported by Raskin and Kircher (2014), two-stage scoring did not produce increased accuracy with the Utah scoring using the Monte Carlo model.

Table 1. Mean (standard deviation) and {95% confidence interval} for criterion accuracy of the four question event-specific test format with three test charts.

	Utah 7-position/GTR .05/.05	Utah 7-position/TSR .05/.05	ESS/TSR .05/.05	ESS/TSR.05/.l0
Unweighted accuracy	.94 (.03) {.89 to .99}	.94 (.03) {.88 to .98}	.94 (.03) {.88 to .98}	.92 (.03) {.86 to .97}
Unweighted inconclusives	.14 (.04) {.07 to .21}	.13 (.03) {.06 to .20}	.16 (.04) {.09 to .23}	.08 (.03) {.03 to .14}
True positive (sensitivity)	.80 (.06) {.69 to .91}	.81 (.06) {.69 to .91}	.81 (.06) {.70 to .91}	.81 (.06) {.69 to .91}
True negative (specificity)	.82 (.05) {.71 to .92}	.82 (.05) {.71 to .92}	.77 (.06) {.65 to .88}	.87 (.05) {.78 to .96}
False negative	.05 (.03) {<.01 to .12}	.05 (.03) {<.01 to .12}	.051 (.03) {<.01 to .12}	.10 (.04) {.02 to .19}
False positive	.05 (.03) {<.01 to .12}	.06 (.03) {<.01 to .13}	.06 (.03) {<.01 to .13}	.05 (.03) {<.01 to .13}
Guilty inconclusive	.15 (.05) {.06 to .26}	.14 (.05) {.05 to .25}	.14 (.05) {.05 to .25}	.09 (.04) {.02 to .19}
Innocent inconclusive	.18 (.05) {.04 to .22}	.12 (.07) {.04 to .22}	.18 (.05) {.08 to .29}	.07 (.04) {<.01 to .15}
PPV	.94 (.04) {.85 to >.99}	.93 (.04) {.85 to >.99}	.93 (.04) {.85 to >.99}	.94 (.04) {.85 to >.99}
NPV	.94 (.04) {.86 to >.99}	.94 (.04) {.86 to >.99}	.94 (.04) {.86 to >.99}	.90 (.05) {.81 to .98}

Table 2 shows the profile of criterion accuracy with Utah 7-position and ESS scores using the GTR and TSR with five test charts. As expected, unweighted accuracy and error rates converged toward the probability cutscores when a sufficient quantity of data is used.

Table 2. Mean (standard deviation) and {95% confidence interval} for Utah 7-position and ESS scores with 5 test charts using grand total and two-stage decision rules with probability cutting scores for deceptive/truthful cutscores (GTR=grand total rule, TSR = two-stage rule).

	Utah 7-position/GTR 5-charts .05/.05	Utah 7-position TSR 5-charts .05/.05	ESS GTR 5-charts .05/.05	ESS TSR 5-charts .05/.05
Unweighted accuracy	.95 (.02) {.91 to .99}	.95 (.02) {.91 to .99}	.95 (.02) {.90 to .99}	.95 (.02) {.90 to .98}
Unweighted inconclusives	.01 (.01) {<.01 to .02}	<.01 (.01) {<.01 to .02}	.02 (.01) {<.01 to .05}	.01 (.01) {<.01 to .04}
True positive (sensitivity)	.94 (.03) {.87 to >.99}	.94 (.03) {.87 to >.99}	.93 (.04) {.85 to >.99}	.94 (.03) {.87 to >.99}
True negative (specificity)	.95 (.03) {.88 to >.99}	.95 (.03) {.88 to >.99}	.93 (.04) {.85 to >.99}	.93 (.04) {.85 to >.99}
False negative	.05 (.03) {<.01 to .12}	.05 (.03) {<.01 to .12}	.05 (.03) {<.01 to .12}	.05 (.03) {<.01 to .12}
False positive	.05 (.03) {<.01 to .11}	.05 (.03) {<.01 to .12}	.05 (.03) {<.01 to .12}	.06 (.03) {<.01 to .12}
Guilty inconclusive	.01 (.01) {<.01 to .04}	<.01 (.01) {<.01 to .02}	.02 (.02) {<.01 to .06}	.01 (.01) {<.01 to .04}
Innocent inconclusive	.01 (.01) {<.01 to .04}	<.01 (.01) {<.01 to .02}	.02 (.02) {<.01 to .07}	.01 (.02) {<.01 to .05}
PPV	.95 (.03) {.89 to >.99}	.95 (.03) {.88 to >.99}	.95 (.03) {.88 to >.99}	.94 (.03) {.87 to >.99}
NPV	.95 (.03) {.88 to >.99}	.95 (.03) {.88 to >.99}	.95 (.03) {.88 to >.99}	.95 (.03) {.88 to >.99}

Table 3 shows the results of an additional analysis conducted to evaluate accuracy of the four-question event-specific test format with traditional cutting scores recommended in previous publications that described the use of this format with the Utah 7-position scoring system. Previous research (Bell, et al., 1999) showed that cutting scores of +6 for truth-telling and -6 for deception provided a high level of decision accuracy with balanced test sensitivity and specificity. The criteria were chosen using performance curves, not probabilities. Comparison of these cutting scores with the referenced distribution in Appendix B shows that they correspond to probability cutting scores of .03 and .03 for deception and truth telling based upon these samples of data.

For the 7-position Utah scoring, unweighted accuracy and other accuracy metrics, including FN and FP errors, were within the range predicted by these cutting scores (.03 and .03). The inconclusive rate for the three-chart data exceeded the APA 20% limit for inconclusive outcomes (2011). However, these results concur with others suggesting that when three-chart totals are inconclusive, conducting additional charts will reduce inconclusive outcomes while maintaining high accuracy (Bell, et al., 1999; Senter & Dollins, 2003; Senter & Dollins, 2004; Senter, Dollins & Krapohl, 2004). Overall accuracy did not change appreciably with the addition of more test charts, but sensitivity and specificity were increased because inconclusive rates were reduced to levels commonly observed in field settings.

Table 3. Four-question event-specific exams Utah 7-position scores with traditional integer cutting scores +6 / - 6 (.03 / .03).

	Three charts	Four charts	Five charts
Unweighted accuracy	.96 (.02) {.91 to >.99}	.97 (.02) {.92 to >.99}	.97 (.02) {.94 to >.99}
Unweighted inconclusives	.23 (.04) {.15 to .32}	.12 (.03) {.06 to .18}	.05 (.02) {.01 to .10}
True positive (sensitivity)	.72 (.08) {.58 to .84}	.84 (.05) {.74 to .94}	.91 (.04) {.83 to .98}
True negative (specificity)	.76 (.06) {.64 to .88}	.86 (.05) {.76 to .96}	.92 (.04) {.84 to .98}
False negative	.03 (.03) {<.01to .09}	.03 (.03) {<.01 to .00}	.03 (.03) {<.01 to .09}
False positive	.04 (.02) {<.01to .08}	.03 (.02) {<.01 to .08}	.03 (.02) {<.01 to .08}
Guilty inconclusive	.25 (.06) {.14 to .39}	.13 (.05) {.04 to .23}	.06 (.03) {<.01 to .13}
Innocent inconclusive	.21 (.06) {.10 to .33}	.11 (.04) {.04 to .20}	.05 (.03) {<.01 to .12}
PPV	.96 (.03) {.88 to >.99}	.97 (.03) {.90 to >.99}	.97 (.03) {.91 to >.99}
NPV	.96 (.03) {.90 to >.99}	.97 (.03) {.91 to >.99}	.97 (.03) {.91 to >.99}

An additional set of analyses was completed to further investigate the effectiveness of the subtotal scoring rules with four-question event-specific examinations for which the RQs are treated as if they were independent. Table 4 shows the profile of criterion accuracy for subtotal scoring rules (SSR) when decisions were

made using Utah 7-position scores with five test charts (there is no reason to suspect this will be much different for ESS scores). Also shown in Table 4 are the results using the SSR with the use of a Bonferroni[4] correction to the probability cutscore for deceptive classifications. As expected, use of the Bonferroni correction substantially reduced false positive errors, corresponding with a reduction in test sensitivity to deception. There was no change to test specificity using the Bonferroni correction. Because multiple statistically significant truthful subtotal scores were required to make a truthful classification, resulting in a deflation of the observed error rate and decrease in test specificity, the p-values for truthful classification were statistically corrected using the Šidák[5] correction. Results using the Šidák correction are shown in the last column of Table 4. As expected, use of this correction did increase test specificity, though mean test specificity remained less than .5.

Effective use of statistical corrections can constrain errors to desired levels with only a marginal improvement in test specificity to truth-telling. Because all questions in the Monte Carlo model had a common criterion state, there was no expectation that test questions varied independently. The potential advantage of the use of subtotal scoring rules was a very low false negative error rate[6]. It is uncertain whether this is represents a practical advantage because polygraph testing is intended to discriminate deception and truth-telling, and an inability identify truth telling due to weak test specificity will result in a biased test format that only weakly discriminates.

Table 5 shows the detection efficiency coefficients produced by the four RQ event specific models shown in Tables 1, 2 and 3. The 95% confidence intervals for all four question MGQT models scored with grand total and two-stage scoring rules included the coefficients for event-specific examinations with three RQs (previously described in this report). Mean detection efficiency coefficients for four question event-specific models scored with grand total and two stage rules ranged from .80, to .91, with the strongest coefficients produced by models with five test charts using decisions based on grand total scores.

[4] Bonferroni correction is applied to a probability cutscore by dividing the desired probability cutscore by the number of statistical classifications. For event-specific polygraph examinations with four RQs the Bonferroni corrected probability cutting score is .05 / 4 = .0125. Inflation of errors due to multiplicity effects will result in an error rate at or near the desired level (.05).

[5] The Šidák correction is used to correct for the increase in inconclusive results that occurs when innocent persons are expected to produce 4 statistically significant truthful scores in order to achieve a truthful test outcome. The Šidák correction is 1-(1-probability cutting score)^1/number of RQs. The inverse of this correction is 1-(1-probability cutting score)^number of RQs can be applied to the probability cutting scores as an alternative to correcting to the p-values.

[6] This is effectively a strategy of achieving a high sensitivity and low false negative errors by using a biased estimator that systematically classifies a larger proportion of persons as deceptive and a smaller proportion of persons as truthful.

Table 4. Criterion accuracy with subtotal scoring rules.

	Utah 7-position SSR no statistical correction	Utah 7-position SSR with Bonferroni correction for deceptive results	Utah 7-position SSR with Bonferroni correction for deceptive results and Šidák correction for truthful results
Unweighted accuracy	.68 (.03) {.62 to .75}	.89 (.04) {.80 to .97}	.90 (.04) {.83 to .99}
Unweighted inconclusives	.03 (.02) {<.01 to .07}	.38 (.05) {.30 to .48}	.35 (.05) {.26 to .44}
True positive (sensitivity)	.99 (.01) {.98 to >.99}	.79 (.06) {.67 to .90}	.79 (.06) {.67 to .90}
True negative (specificity)	.35 (.07) {.22 to .48}	.35 (.07) {.22 to .48}	.41 (.07) {.28 to .55}
False negative	<.01 (<.01) {<.01 to .02}	<.01 (<.01) {<.01 to .02}	<.01 (<.01) {<.01 to .02}
False positive	.59 (.07) {.45 to .72}	.10 (.04) {.02 to .18}	.10 (.04) {.02 to .18}
Guilty inconclusive	<.01 (<.01) {<.01 to .021}	.21 (.06) {.10 to .33}	.21 (.06) {.10 to .33}
Innocent inconclusive	.06 (.04) {<.01 to .14}	.56 (.07) {.42 to .69}	.50 (.07) {.35 to .62}
PPV	.63 (.06) {.51 to .73}	.89 (.05) {.80 to .98}	.89 (.05) {.80 to .98}
NPV	>.99 (.01) {.94 to >.99}	>.99 (.01) {.94 to >.99}	>.99 (.02) {.95 to >.99}

Table 5. Monte Carlo estimates of detection efficiency coefficients (standard errors) and 95% confidence intervals for event-specific exams with four relevant questions.

Scoring method	Decision rule	Number of test charts	Probability cutscores (deception / truth)	Mean detection efficiency coefficient (SE) {95% CI}
Utah 7-position	GTR	3 charts	.05 / .05	.82 (.05) {.71 to .91}
Utah 7-position	TSR	3 charts	.05 / .05	.82 (.05) {.71 to .91}
ESS	TSR	3 charts	.05 / .05	.80 (.05) {.69 to .89}
ESS	TSR	3 charts	.05 / .10	.80 (.06) {.68 to .90}
Utah 7-position	GTR	3 charts	-6 (.029) +6 (.029)	.81 (.05) {.71 to .88}
Utah 7-position	GTR	4 charts	-6 (.029) +6 (.029)	.88 (.04) {.79 to .94}
Utah 7-position	GTR	5 charts	-6 (.029) +6 (.029)	.91 (.04) {.84 to .97}
Utah 7-position	GTR	5 charts	.05 / .05	.90 (.04) {.81 to .98}
Utah 7-position	TSR	5 charts	.05 / .05	.90 (.04) {.81 to .97}
ESS	GTR	5 charts	.05 / .05	.89 (.04) {.79 to .96}
ESS	TSR	5 charts	.05 / .05	.89 (.05) {.79 to .96}
Utah 7-position	SSR	5 charts	.05 / .05 uncorrected	.49 (.06) {.38 to .61}
Utah 7-position	SSR	5 charts	.05 Bonferroni corrected / .05	.70 (.05) {.60 to .80}
Utah 7-position	SSR	5 chart	.05 Bonferroni / .05 Šidák	.72 (.05) {.61 to .81}

Detection efficiency statistics using subtotal scoring rules were weaker and the upper limit of the 95% confidence interval was less than values previously reported for event-specific exams, regardless of the number of test charts and regardless of the use of statistical corrections to constrain errors and inconclusive results.

To further contrast these results, we also calculated the detection efficiency for AFMGQT models reported by APA (2011). These included 7-position scores for two to four RQs and 3 test charts when scored with subtotal scoring rules and traditional cutscores (r = .59) and ESS scores for two to four RQs and 3 test charts when scored with subtotal scoring rules with Šidák correction for truthful classifications with probability cutscores set at .05 and .10 for deceptive and truthful classifications (r = .66).

Discussion

This project involved the use of two archival samples, one laboratory and one field study. We used two different, though related methods for manual test data analysis to develop statistical reference distributions and calculate the range of out of sample error (i.e., generalization error that can be expected to be observed in other sample results or field experience) in the form of 95% confidence intervals. These confidence intervals show the upper and lower limit of expected accuracy for event-specific polygraph examinations with a combination of four primary and secondary RQs. The archival data were from two samples of event-specific examinations with a combination of three primary and secondary RQs. The grand-total mean and variance of the four-question format was calculated through non-parametric bootstrap resampling of the sub-total scores of the primary and secondary subtotal scores of the sampling data. Grand total mean and variance estimates for the four-question format were then used as seed data for a parametric bootstrap that calculated the 95% confidence intervals t used to calculate the expected range of out-sample generalization error.

This approach depends on the assumption that subtotal mean and variance may be similar for event-specific examinations with three and four RQs and that information and knowledge may be generalized from one format to the other. The general assumption underlying this approach is that responses for innocent examinees will be loaded more onto CQs while responses for guilty examinees will be loaded more onto RQs, regardless of the number of questions. There is no indication in the published literature and no plausible hypothetical rationale to suggest that the direction of differential salience of the test questions changes as a function of the number of RQs. Although differences in the linear magnitude of differential responses to CQs and RQs has not yet been studied, scoring systems that do not make use of linear assumptions will be unaffected by those differences, as long as the direction of differential responses to CQs and RQs varies similarly for guilty and innocent examinees regardless of the number of RQs. It will be important to continue to evaluate the proposed reference distributions in field and laboratory studies.

Results of this Monte Carlo analysis showed that event-specific examinations with four RQs can provide high rates of criterion accuracy with both guilty and innocent examinees when using grand total and two stage decision rules with Utah 7-position scores and ESS scores. Inspection of the 95% confidence intervals indicated that unweighted accuracy, inconclusive rates, sensitivity, specificity, false positives, false negatives, and positive and negative predictive values (shown in Table 1) were similar for the two scoring methods. Unweighted mean accuracy was in excess of 90%, which exceeded the previously reported mean accuracy for AFMGQT formats scored with subtotal scoring rules. Unweighted inconclusive rates were well under the 20% ceiling (APA, 2012) and under that reported for formats scored with subtotal scores.

Test sensitivity to deception using grand total and two-stage rules consistently equaled or exceeded that reported for test sensitivity using subtotal scoring rules under all conditions except when the cutscores and decision rules were biased heavily against truth-tellers. Mean unweighted accuracy and error rates with five test charts (shown in

Table 2) were at or near the rates predicted by the prior selected probability cutscores, suggesting that a sufficient quantity of test data could allow informed judgments regarding the meaning and utility of the test results.

The only potential advantage to the use of subtotal scoring rules was a very low false negative error rate and a very high sensitivity rate (see Table 4). However, this occurred at the expense of a high false positive error rate (.59), mean test specificity levels that remained less than .5, and substantially less than that for decisions using the grand total and two-stage rules. The lower limit of the 95% confidence interval for test specificity exceeded the chance probability of .063[7] for four RQs with subtotal scoring, though we question the use of a test based on this chance probability. Along with weak test specificity, there was a substantial increase in false-positive errors for the subtotal scoring rules. Use of common statistical corrections reduced the false positive and false negative error rates, but they also reduced mean test sensitivity to a level that was substantially less than that achieved by the grand total decision rule. Mean inconclusive rates exceeded the 20% ceiling requirement of the APA (2011) with the subtotal scoring rules. Completion of up to five test charts did not adequately correct these deficiencies.

We suggest that a more effective means to achieve a very low false negative error rate would use the grand total or two-stage decision rules with the probability cutscore selected to achieve a desired tolerance for error. An example is shown in Table 6, for which the probability cutscore for truth telling was set to achieve a false-negative error rate near .01. Grand total decisions can provide desirable sensitivity levels in a manner that also constrains inconclusive results and false-positive errors to desired tolerances, while also providing test specificity at rates for which the lower limit of the 95% confidence interval that easily exceeds the .5 level.

[7] Calculated as 0.5^4 because subtotal scoring rules with 4 RQs require 4 truthful question results in order to achieve a truthful test result.

Table 6. Criterion accuracy using grand total decisions with 5 charts (.05 / .01).

	Utah 7-position GTR .05 / .01	ESS GTR .05 / .01
Unweighted accuracy	.97 (.02) {.93 to >.99}	.97 (.02) {.93 to >.99}
Unweighted inconclusives	.08 (.03) {.03 to .13}	.08 (.03) {.03 to .13}
True positive (sensitivity)	.94 (.03) {.87 to >.99}	.94 (.03) {.87 to >.99}
True negative (specificity)	.85 (.05) {.74 to .94}	.85 (.05) {.74 to .94}
False negative	.01 (.01) {<.01 to .04}	.01 (.01) {<.01 to .04}
False positive	.05 (.03) {<.01 to .11}	.05 (.03) {<.01 to .11}
Guilty inconclusive	.05 (.03) {<.01 to .12}	.05 (.03) {<.01 to .12}
Innocent inconclusive	.11 (.04) {.02 to .2}	.11 (.04) {.02 to .2}
PPV	.95 (.03) {.89 to >.99}	.95 (.03) {.89 to >.99}
NPV	.99 (.02) {.95 to >.99}	.99 (.02) {.95 to >.99}
Detection efficiency coefficient	.90 (.04) {.83 to .96}	.90 (.04) {.83 to .96}

Limitations

Monte Carlo modeling techniques depend on assumptions about the representativeness of the seed data and the adequacy of our knowledge about interaction among the variables. The estimated reference distributions (shown in Appendices A and B) are statistical approximations of the population distributions of four-question event-specific examinations of guilty and innocent persons. These reference distributions were computationally imputed from the data of other examination formats with three RQs. Therefore, their appropriateness depends on an assumption of similarity of the differences in response magnitude for RQs and CQs, regardless of the number of RQs included in the test format. Although there is no plausible hypothesis or rationale to suggest expected differences in the general pattern of responses as a function of the number of questions[8], this assumption has not been thoroughly investigated due to the lack of previously available data for event specific exams with four RQs. Another limitation to Monte Carlo methods, which is shared by other methods, is that reference distributions are empirical and their generalizability is limited by the data upon which they were based. Access to a distribution of sampling distributions may answer questions about the degree to which the proposed reference distributions converge towards the population distributions. The present analysis serves is an initial inquiry into the use of statistical reference distributions as a basis for the selection of cutscores to achieve desired test accuracy and error rates with four-question event-specific exams.

Use of a parametric bootstrap for the Monte Carlo model represents an explicit assumption that the population distributions of guilty and innocent scores can be characterized by a Gaussian distribution. Non-parametric bootstrapping would not require this assumption, but would provide less general information about the application of Gaussian-Gaussian, or equivariance-Gaussian signal discrimination models to polygraph examinations. Explicit reliance on an assumed normal distribution simplifies assumptions about generalization of the Gaussian model at the expense of providing less information about the performance characteristics of the seed data. The choice of a parametric bootstrap provided general information about the testing model that would be less affected by any idiosyncrasies of the seed data. Replication of this analysis with non-parametric bootstrapping and samples of both laboratory and field cases is needed.

There is also a need for more information regarding the differences between ESS and Utah 7-position scores and the observed interaction effects with the seed data for the ESS scores. Seed data for the ESS scores showed that the first RQ produced more positive scores only for innocent persons, while the Utah 7-position scores showed that the first RQ produced a more positive score for both innocent and guilty persons. It is possible that the observed difference was be due to sampling differences or differences in scoring skills among the evaluators.

Another possible source of the aforementioned interaction is the difference between the question formats used to collect the Utah and ESS data. ESS scores were obtained by scoring the first RQ to the stronger of the preceding or subsequent CQ, whereas the Utah scores were obtained by scoring all RQs to the preceding CQ. This procedure is more likely to affect the scores of R1 for innocent cases.

[8] Guilty persons are expected to produce generally larger responses to RQs than CQs regardless of the number of questions, while the opposite pattern is expected from innocent persons.

Regardless of the cause of the interaction and main effect for the first RQ, an important practical implication of this finding is that rotation of RQs may be important to counterbalance the position effects across three charts.

Another important finding was the lack difference between the R2 and R3 questions in either data set. It is important to note that the traditionally hypothesized effect for differences in the scores of the third RQ was not confirmed with either the Utah 7-position or ESS scores. This questions previously held assumptions about response variance based on the semantics of the stimuli. Another limitation of the Monte Carlo model is that these models did not take into account the possibility that habituation might occur with a four-RQ test. Although it may be possible to develop a Monte Carlo model that accounts for within-test habituation, more information will be needed to characterize this effect.

We did not investigate why two-stage decision rules did not optimize test performance as reported in other studies; this may be a limitation of the Monte Carlo design. It is possible that the advantages two-stage decision rules are reduced when test performance is optimized by a larger number of stimulus presentations and the completion of up to five test charts. Further research is needed in this area.

We did not make a statistical comparison of the effectiveness of decision cutscores based on Gaussian assumptions with cutscores based on performance curves or heuristic experience. Results from previous studies and this analysis suggest that both methods can be effective. This may be an important area for further study to better understand the potential advantages of cutscores based on performance characteristi cs and distributional characteristics.

Finally, some discussion is in order regarding the use of laboratory and field samples. Seed data for the Utah 7-position distributions were from a laboratory study, while the seed data for the ESS distributions were from field cases. Although the two approaches can have different advantages, field and laboratory studies have shown similar results (APA, 2011; Office of Technology Assessment, 1983; National Research Council, 2003; Pollina et al., 2004; Raskin & Honts, 2002). High correspondence between field and laboratory results is consistent with evidence in other fields of study (Anderson, Lindsay, & Bushman, 1999).

Data from field studies provide the potential for greater ecological validity and greater generalizability, though this advantage may be reduced by complications from systematic sampling bias that occurs as a result of non-random confirmation of ground truth criteria and limited ability to completely control potentially confounding variables. Laboratory studies typically employ random selection of examinees to guilty and innocent groups so that results of laboratory studies may be more generalizable to other sampling data. In addition, laboratory studies offer the potential for sufficient control over confounding variables such that inferences may be made about causality. The most effective approach is to use all types of studies to strategically advance our knowledge, and the results of this Monte Carlo analysis[9]– provide additional support for the similarity of the general pattern of results and effectiveness from laboratory and field studies.

Conclusion

This study provides general support for the effectiveness of event-specific polygraph examinations with four RQs and for the effectiveness of decision cutscores selected using both performance curves and statistical distributions.

[9]Monte Carlo studies offer another type of investigation strategy, centered on the study of the decision problem itself while making assumptions about the generalizability of our prior knowledge.

Mean detection efficiency coefficients for event-specific exams with four RQs using grand-total and two-stage decision rules equaled or exceeded rates previously reported for event-specific exams with three RQs, indicating a potential advantage of four RQs in event specific exams. Unweighted accuracy was .92 to .94 for three test charts, and converged towards the .95 confidence level predicted by the probability cutscores (.05 / .05) when five test charts were used.

As a practical matter, event-specific examinations with four RQs can provide perceived or practical advantages to field examiners, referring professionals, and agencies that use polygraph examination results. These advantages include a potential increase in precision, and the ability to use a wider array of questions to provide a sense of satisfaction to examiners and referring professionals. However, it is important to recognize that the RQs do not vary independently and one cannot make accurate decisions based on scoring of individual RQs. When decisions are based on total scores with the Utah 7-position scoring and ESS with the RQs, these examinations produce balanced sensitivity, specificity, false positive, and false negative errors. Use of four RQs can lead to increases in the internal consistency of the CQT, although we did not test for that effect.

This analysis showed no scientific and no practical advantage to the use of subtotal scores as a basis for decisions for event-specific examinations with four RQs, findings that concur with previous studies. These studies have failed to demonstrate that test questions vary independently within event-specific exams and that neither test sensitivity nor test specificity can be optimized by making decisions at the level of the individual questions. Decisions using grand total and two-stage decision rules, coupled with a scientific approaches to the selection of decision cutscores, can more effectively achieve low false negative and false positive errors, high criterion accuracy, and low inconclusive rates. Use of the independence hypothesis and associated decision rules based on subtotal scores was associated with less accurate test results for the event-specific cases in this analysis. Decisions using grand total scores are less prone to systematic bias and over-classification of deception than those based on subtotal scores, which are more prone to poor test performance with innocent persons.

These results provide further contradiction to hypotheses or assumptions that RQs of event-specific polygraph examinations vary independently when they attempt to employ different action verbs or attempt to describe different behavioral aspects or different levels of involvement in a known or alleged incident or context. This is not surprising, as no previous study has supported the hypothesis of independent variance or expectations of improved test performance as a function of attempts to interpret results at the level of question subtotal scores.

We also note that assumptions about independent response variance are inconsistent with scientific and statistical theory, for which independence assumes no sources of shared response variance. In actuality, all RQs within an investigation of a known incident can be expected to have potential sources of shared response variance, beginning with the examinee. This has been shown in previous study results that failed to support the hypothesis that responses to different RQs vary independently (Barland, Honts & Barger, 1989; Krapohl & Norris, 2000; Podlesney & Truslow, 1993; Raskin, Kircher, Honts & Horowitz, 1988; Senter, 2003; Senter & Dollins, 2003). Consistent with these studies, the present study, found that grand total and two-stage decision rules outperformed the criterion accuracy rates achieved by subtotal scoring rules in.

It would be a mistake to believe that our present knowledge is adequate or complete regarding the precise distributions of scores for guilty and innocent persons, and we suggest that there is a need to quantify and document our knowledge regarding these distributions. Neglecting to document the characteristics of available sampling distributions can contribute to systematic and deliberate reliance on cutscores and decision rules that are not supported by scientific evidence. Intransigent reliance on

traditional approaches that are inconsistent with published scientific evidence weakens the scientific credibility of the profession and hinders progress towards the development and implementation of improved polygraph techniques. Increased reliance on scientific methods and scientific evidence can enable both examiners and referring professionals to make better informed selection of testing techniques, including cutting scores and decision rules based on statistical estimates of expected performance or error rates, along with consideration for a determined tolerance or payoff matrix for false positive and/or false negative errors.

We recommend continued interest and additional research in event-specific examinations with four RQs using grand total and two-stage decision rules. We realize this approach differs slightly from the traditional approach, and we remind readers that it is important not to cling to traditional approaches for no reason other than tradition, especially when traditional approaches are inconsistent with scientific evidence.

Finally, it is important to emphasize greater attention to the lower limits of the confidence intervals for accuracy and the upper limits for errors and inconclusive results. The limits of the confidence intervals provide more useful information towards estimation of the worst case scenario than the reported mean estimates. Confidence intervals in this analysis provide a more reasonable and cautious estimate of out of sample generalization error for the testing model. Considering worst-case boundaries will provide consumers with the conservative range of potential performance. This analysis suggests that the lower limits of test performance for event-specific examinations with four RQs can provide desirable levels of criterion accuracy that equal or exceed that of other validated polygraph techniques. The distributional data from this study allows examiners to use the four-RQ approach in an event-specific setting and provide consumers with error estimates within the limitations of the datasets sampled. Grand total scoring of event-specific examinations with four RQs appears to offers high criterion accuracy along with more desirable inconclusive and error rates than subtotal scoring, while constraining errors to specified tolerance rates through the use of thoughtfully selected cutting scores.

References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In N.J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Sage.

Abdi, H. (2010). The Greenhouse-Geisser correction. In Neil Salkind (Ed.), *Encyclopedia of Research Design.* Sage. American Polygraph Association (2011). Meta-analytic survey of criterion accuracy of validated polygraph techniques. *Polygraph*, *40*(4).

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions In Psychological Science, 8,* 3-9.

Barland, G. H., Honts, C. R., & Barger, S. D. (1989). *Studies of the accuracy of security screening polygraph examinations.* Department of Defense Polygraph Institute.

Bell, B. G., Raskin, D. C., Honts, C. R., & Kircher, J. C. (1999). The Utah numerical scoring system. *Polygraph, 28*, 1-9.

Carsey, T. M. & Harden, J. J. (2014). *Monte Carlo Methods and Resampling Methods for Social Science.* Sage.

Department of Defense (2006a). *Federal Psychophysiological Detection of Deception Examiner Handbook. Retrieved* from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02- 10-2006.pdf on 3-31-2007. Reprinted in *Polygraph*, *40*(1), 2-66.

Department of Defense (2006b). *Psychophysiological Detection of Deception Analysis II -- Course #503. Test data analysis: DoDPI numerical evaluation scoring system.* Available from the author. (Retrieved from http://www.antipolygraph.org/documents/federal-polygraph-handbook-02-10-2006.pdf on 3-31-2007).

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika 68* (3): 589–599.

Girden, E. (1992). *ANOVA: Repeated measures.* Newbury Park, CA: Sag

Kircher, J. C., Horowitz, S. W. & Raskin, D.C. (1988). Meta-analysis of mock crime studies of the control question polygraph technique. *Law and Human Behavior, 12,* 79-90.

Kircher, J. C., & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. *Journal of Applied Psychology*, *73*, 291-302.

Krapohl, D. J. & Norris, W. F. (2000). An exploratory study of traditional and objective scoring systems with MGQT field cases. *Polygraph, 29,* 185-194.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics 11 (2),* 204–209.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science. Special Issue,* 125-130.

National Research Council (2003). *The Polygraph and Lie Detection.* Committee to Review the Scientific Evidence on the Polygraph. Washington, DC: National Academies Press.

Nelson, R., Krapohl, D., Handler, M., 2008. Brute force comparison: a Monte Carlo study of the objective scoring system version 3 (OSS-3) and human polygraph scorers. *Polygraph* 37 (3), 185-215.

Nelson, R. Handler, M. Shaw, P., Gougler, M., Blalock, B., Russell, C., Cushman, B., & Oelrich, M. (2011). Using the Empirical Scoring System, *Polygraph*, *40*(2).

Office of Technology Assessment (1983). *The validity of polygraph testing: A research review and evaluation.* Re-printed in *Polygraph*, *12*, 198-319

Podlesny, J. A., Truslow, C. M., 1993. Validity of an expanded-issue (modified general question) polygraph technique in a simulated distributed-crime-roles context. *Journal of Applied Psychology, 78,* 788-797.

Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J. & Ryan, A. H. (2004). Comparison of polygraph data obtained from individuals involved in mock crimes and actual criminal investigations. *Journal of applied psychology, 89,* 1099-105.

Raskin, D. C. & Hare, R.D. (1978). Psychopathy and detection of deception in a prison population. *Psychophysiology, 15,* 126-136.

Raskin, D.C. & Honts, C.R. (2002). The comparison question test. In M. Kleiner (Ed.), *Handbook of polygraph testing*. London: Academic (1-49).

Raskin, D. C. & Kircher, J. C. (2014). Validity of polygraph techniques and decision methods. In D. C. Raskin, C. R. Honts, & J. C. Kircher (Eds.), *Credibility Assessment: Scientific Research and Applications*. San Diego, CA: Elsevier/Academic Press.

Raskin, D.C., Kircher, J.C., Honts, C.R. and Horowitz, S.W. (1988) *A Study of Validity of Polygraph Examinations in Criminal Investigations*. Grant number 85-IJ-CX-0040. Salt Lake City: Department of Psychology, University of Utah.

Rovner, L. I., Raskin, D. C. & Kircher, J.C. (1979). Effects of information and practice on detection of deception. *Psychophysiology, 16,* 197-198 (abstract).

Senter, S. (2003). Modified General Question Test Decision Rule Exploration. *Polygraph 32*(4) 251-263.

Senter, S. & Dollins, A. (2004). Comparison of question series and decision rules: A replication. *Polygraph, 33,* 223-233.

Senter, S., Dollins, A. & Krapohl, D. (2004). A Comparison of Polygraph Data Evaluation Conventions Used at the University of Utah and the Department of Defense Polygraph Institute. *Polygraph, 33,* 214-222.

Silverman, B. W. and Young, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika 74(3).*

Appendix A.

AFMGQT with two to four relevant questions with probability cutting scores = .05/.10 for deception and truth-telling (APA, 2011) and traditional cutting scores of +/-3.

	AFMGQT/ESS Sub-total scoring rules (.05/.10)	AFMGQT/7-position Sub-total scoring rules (-3/+3)
Unweighted accuracy	.875 (.039) {.798 to .953}	.817 (.042) {.734 to .900}
Unweighted inconclusives	.170 (.036) {.100 to .241}	.197 (.030) {.138 to .255}
True positive (sensitivity)	.729 (.065) {.603 to .856}	.783 (.058) {.669 to .896}
True negative (specificity)	.700 (.063) {.577 to .823}	.538 (.068) {.405 to .672}
False negative	.092 (.046) {.002 to .182}	.079 (.050) {.001 to .177}
False positive	.112 (.047) {.020 to .204}	.203 (.057) {.090 to .315}
Guilty inconclusive	.178 (.056) {.068 to .289}	.137 (.033) {.071 to .202}
Innocent inconclusive	.162 (.047) {.071 to .254}	.257 (.049) {.160 to .354}
PPV	.864 (.058) {.751 to .977}	.79 (.059) {.675 to .905}
NPV	.887 (.052) {.785 to .989}	.874 (.062) {.753 to .996}

Appendix B.

Reference distributions for Utah 7-position grand total scores of four RQ event-specific exams.
Mean deceptive 7-position score = -11 (SD = 9) Mean truthful 7-position score = +13 (SD = 10)
Note: The probability values given here are valid only for the data set from which they were calculated. Generalizability to other data is unknown.

Truthful		Deceptive	
CutScore	p-value	CutScore	p-value
1	.091	0	.097
2	.074	-1	.081
3	.060	-2	.067
4	.048	-3	.055
5	.038	-4	.045
6	.029	-5	.036
7	.023	-6	.029
8	.017	-7	.023
9	.013	-8	.018
10	.010	-9	.014
11	.007	-10	.011
12	.005	-11	.008
13	.004	-12	.006
14	.003	-13	.005
15	.002	-14	.003
16	.001	-15	.003
17	.001	-16	.002
18	.001	-17	.001
19	<.001	-18	.001
		-19	.001
		-20	<.001

Appendix C.

Reference distributions for ESS grand total scores of four-RQ event-specific exams.
Mean deceptive ESS score = -12 (SD = 10) Mean truthful ESS score = +11 (SD = 9)
Note: The probability values given here are valid only for the data set from which they were calculated. Generalizability to other data is unknown.

Truthful Scores		Deceptive Scores	
CutScore	p-value	CutScore	p-value
1	.097	0	.111
2	.081	-1	.091
3	.067	-2	.074
4	.055	-3	.060
5	.045	-4	.048
6	.036	-5	.038
7	.029	-6	.029
8	.023	-7	.023
9	.018	-8	.017
10	.014	-9	.013
11	.011	-10	.010
12	.008	-11	.007
13	.006	-12	.005
14	.005	-13	.004
15	.003	-14	.003
16	.003	-15	.002
17	.002	-16	.001
18	.001	-17	.001
19	.001	-18	.001
20	.001	-19	<.001
21	<.001		

Appendix D. Decision rules.

Grand Total Rule:

A. If the grand total >= NDI cutting score, then NDI

B. If the grand total <= DI cutting score, then DI

C. All other results are inconclusive

Grand total decisions provide the highest overall classification accuracy.

Sub-total Scoring Rules:

A. If *all* sub-totals >= NDI or NSR sub-total cutting score, then NDI or NSR

B. If *any* sub-total <= DI or SR sub-total cutting score, then DI or SR

C. All other results are inconclusive

Sub-total decisions attempt to interpret independent response variance.

Two-Stage Rules (Senter 2003; Senter & Dollins, 2008)

Stage 1: Grand total rule (do not use the sub-total scores at Stage 1)

A. If the grand total >= NDI cutting score, then NDI

B. If the grand total <= DI cutting score, then DI

Stage 2: Sub-total score rule (only if the grand total is inconclusive at Stage 1)

A. If *any* sub-total <= statistically corrected DI sub-total cutting score , then DI

B. There are no NDI considerations using sub-totals at Stage 2 because these would have been resolved at Stage 1

C. All other results are inconclusive

Two-stage Rules provide decision accuracy similar to the grand total, with a potential increase in test sensitivity and potential decrease in inconclusive results.